CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Nature Precedings

PROJECT SUMMARY

Foregut microbiome in development of esophageal adenocarcinoma

Liying Yang, William E. Oberdorf, Erika Gerz, Tamasha Parsons, Pinak Shah, Sukhleen Bedi, Carlos W. Nossa, Stuart M. Brown, Yu Chen, Mengling Liu, Michael Poles, Fritz Francois, Morris Traube, Navjeet Singh, Todd Z. DeSantis, Gary L. Andersen, Monika Bihan, Les Foster, Aaron Tenney, Daniel Brami, Mathangi Thiagarajan, Indresh Singh, Manolito Torralba, Shibu Yooseph, Yu-Hui Rogers, Eoin L. Brodie, Karen E. Nelson, and Zhiheng Pei.

New York University School of Medicine, Lawrence Berkeley National Laboratory, The J. Craig Venter Institute.

Nature Precedings : doi:10.1038/npre.2010.5026.1 : Posted 18 Oct 2010

I. PROJECT ID NUMBER, PUBLICATION MORATORIUM INFORMATION, PROJECT DESCRIPTION:

This manuscript is part of a pilot effort on the part of NIH staff and the Nature publishing group to provide a more convenient archive for "marker papers" to be published. These "marker papers" are designed to provide the users of community resource data sets with information regarding the status and scope of individual community resource projects. For further information see editorial in September 2010 edition of Nature Genetics (*Nature Genetics,* **42**, 729 (2010)), and the Nature Precedings HMP summary page.

Project ID: 46307.
Length of your publication moratorium: 12 months.

Esophageal adenocarcinoma (EA), the type of cancer linked to heartburn due to gastroesophageal reflux diseases (GERD), has increased six fold in the past 30 years. This cannot currently be explained by the usual environmental or by host genetic factors. EA is the end result of a sequence of GERD-related diseases, preceded by reflux esophagitis (RE) and Barrett's esophagus (BE). Preliminary studies by Pei and colleagues at NYU on elderly male veterans identified two types of microbiotas in the esophagus. Patients who carry the type II microbiota are >15 fold likely to have esophagitis and BE than those harboring the type I microbiota. In a small scale study, we also found that 3 of 3 cases of EA harbored the type II biota. The findings have opened a new approach to understanding the recent surge in the incidence of EA.

Our **long-term goal** is to identify the cause of GERD sequence. The **hypothesis** to be tested is that changes in the foregut microbiome are associated with EA and its precursors, RE and BE in GERD sequence. We will conduct a case control study to demonstrate the microbiome disease association in every stage of GERD sequence, as well as analyze the trend in changes in the microbiome along disease progression toward EA, by two specific aims. **Aim 1** is to conduct a comprehensive population survey of the foregut microbiome and demonstrate its association with GERD sequence. Furthermore, spatial relationship between the esophageal microbiota and upstream (mouth) and downstream (stomach) foregut microbiotas as well as temporal stability of the microbiome-disease association will also be examined. **Aim 2** is to define the distal esophageal metagenome and demonstrate its association with GERD sequence. Detailed analyses will include pathway-disease and gene-disease associations. Archaea, fungi and

viruses, if identified, also will be correlated with the diseases. A significant association between the foregut microbiome and GERD sequence, if demonstrated, will be the first step for eventually testing whether an abnormal microbiome is required for the development of the sequence of phenotypic changes toward EA. If EA and its precursors represent a microecological disease, treating the cause of GERD might become possible, for example, by normalizing the microbiota through use of antibiotics, probiotics, or prebiotics. Causative therapy of GERD could prevent its progression and reverse the current trend of increasing incidence of EA.

II. DATA QUALITY:

Our data quality metrics will continue to follow the guidelines that are under review for the larger Human Microbiome Project (HMP). We will measure the total number of reads and bases generated as well as the count of base calls surpassing a QV score of 20 within the clear range trimmed by the Roche-provided 454 base calling algorithm software.

III. DATA ANALYSIS AND PUBLICATION PLANS:

Data Analysis: All data will be used to analyze the relationship between the normal controls and subjects with RE, BE, and EA. Temporal stability of the esophageal microbiome and the special relationship among the three major foregut sites, mouth, esophagus, and stomach will also be evaluated. Both 16S rRNA gene surveys and metagenomic shotgun sequencing will be conducted on all samples.

Initial sequence analysis includes deconvolution of the data files in sff format which are transformed into untrimmed fasta files (containing base calls) and qual files (containing quality value, QV, scores for each base call) using Roche's software, sffinfo. The positions of the clear range of each read will be recorded. For 16S data, amplification primers and barcodes within each sequence will be located with BLAST using parameters "W 5 -G 2 -q -1 -F F". Reads from the fasta file will be sorted into the individual library corresponding to the barcode producing the highest BLAST score. The positions of barcodes and primers within the untrimmed sequence will be recorded. For metagenomic data, the reads will be sorted into individual libraries based on the barcode information they contain.

The 16S rRNA gene segment will be determined by NAST (DeSantis, 2006) comparison to a reference template set:
(http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/core_set_aligned.fasta.imputed).
Prior to taxonomic classification, base calls with QV scores under 10 will be transformed to N. An ergatis-based compute grid workflow (http://ergatis.sourceforge.net) will be used to classify all reads with at least 42 remaining non-N 8-mers using RDP's Naive Bayesian Classifier version 2.2 (Wang, 2007) and training set No. 6 posted March 2010 (http://rdp-classifier.sourceforge.net). Domain, Phylum, Class, Order, Family and Genus level groupings will be recorded with corresponding bootstrap values. Trimmed sequences will be clustered at 2% sequence divergence to create Operational Taxonomic Units and frequency tables of reads per OTU per sample will be used for beta-diversity analysis.

Nature Precedings : doi:10.1038/npre.2010.5026.1 : Posted 18 Oct 2010

Metagenomic shotgun sequence data will be processed using JCVI's metagenomic annotation pipeline. The metagenomic annotation pipeline identifies and assigns functions to the protein coding genes in a given input sample. This functional annotation will be used to identify protein families and pathways represented in the metagenomic data, and this will subsequently be used to infer the metabolic capabilities of the microbial communities in the samples. We will also use recruit metagenomic reads to reference genomes to identify the distribution of sequenced HMP relevant reference genomes in these samples and address questions of genome structure and variation in normal versus disease samples.

Differentially abundant taxonomic groups, protein families and pathways across samples will be identified using count statistics. In addition, samples will be compared and clustered based on their genetic distances, community structures, and functional profiles, and we will attempt to associate the sample clusters with the collected sample metadata.

Publication Plans: We hope to submit findings from these analyses for publication by the Publication Moratorium date.

IV. DATA RELEASE PLAN:

In keeping with NIH data release policy, all data generated through this project will be made available to the biomedical research community. Consistent with guidelines, the sharing of research data will include the rapid dissemination of research data to the scientific community in order to maximize the public benefit of the data produced. Data sharing will involve a website (http://gerd.med.nyu.edu/) for sharing research results among project investigators. A public portion of the website will have links to published papers through PubMed. Besides publication, we will make our results available to the community of scientists interested in GERD by presentations at national meetings, such as the annual meetings of American Society for Microbiology and Digestive Disease Week. Especially, we will disseminate our findings on the association between GERD and the foregut microbiome to avoid unintentional duplication of research. Furthermore, we would welcome collaboration with others who could make use of the protocols developed in this project.

In accordance with the updated resource sharing policy (http://nihroadmap.nih.gov/hmp/datareleaseguidelines.asp) our data release will be as follows:

I. All 454 sequence data inclusive of 16S rDNA and metagenomic data will be released as quickly as possible (on a weekly basis), to the NCBI. Any metagenomic assemblies and the associated annotations will also be deposited at the NCBI and to GenBank within 45 days in agreement with the proposed policy. The raw traces will also be deposited at the Trace Archive or to the Short Read Archive at NCBI/NLM/NIH. The release of the UH2 dataset to dbGaP and to SRA occurred on July 1, 2010.

II. Wherever possible, we will work closely with the HMP Data Analysis and Coordination Center (DACC) to coordinate standardization, and data release.

III. Metadata will also be submitted to public repositories, and any human sequence or clinical data that could potentially be used as an identifier will be submitted to NCBI's dbGaP (a controlled access database). These data are screened for human genome sequence by NCBI prior to deposition into the open access portion of SRA.

Nature Precedings : doi:10.1038/npre.2010.5026.1 : Posted 18 Oct 2010

IV. Any analyses that are performed on the datasets that are generated through this project will be made available to the public via the DACC as soon as our manuscripts are accepted for publication.

V. CONTACT PERSON:

Dr. Zhiheng Pei, New York University School of Medicine, 212-951-5492, zhiheng.pei@med.nyu.edu.

Nature Precedings : doi:10.1038/npre.2010.5026.1 : Posted 18 Oct 2010